Improving OCR of Black Letter in Historical Newspapers: The Unreasonable Effectiveness of HTR Models on Low-Resolution Images

Ströbel, Phillip ; Clematide, Simon

DH2019
UTRECHT UNIVERSITY 8-12 JULY

Utrecht University

Improving OCR of Black Letter in Historical Newspapers: The Unreasonable Effectiveness of HTR Models on Low-Resolution Images

Phillip Benjamin Ströbel (pstroebel@cl.uzh.ch), University of Zurich, Switzerland and Simon Clematide (simon.clematide@cl.uzh.ch), University of Zurich, Switzerland
XML

1. Introduction

The quality of Optical Character Recognition (OCR) is a decisive factor for the application of text mining techniques on historical newspapers (Chiron et al., 2017; Walker et al., 2010; Strange et al., 2014). OCR for texts published in black letter is particularly challenging due to several factors: the low distinctiveness of characters, the change over time regarding vocabulary and spelling, the use of small font sizes, and the oftentimes poor paper quality.

Holley (2009) argued that in light of the poor OCR quality in newspapers, a focus on manual crowd-correction is more promising than investments in software development. Although automatic OCR post-correction can improve the quality of the text, the methods often lack precision, are not robust enough, or require a lot of in-domain training data (Alex et al., 2012; Chiron et al., 2017).

The problems are manifold and complex, but recent progress in neural OCR techniques promises significant improvements (Springmann and Lüdeling, 2016). These OCR models often outperform commercial systems like **ABBYY FineReader** [1]. However, the training of a neural system using open-source software (e.g., **Tesseract** [2]) is demanding. Integrated handwritten text recognition and annotation platforms like **Transkribus** [3] facilitate the creation of a ground truth, as well as the training and application of neural and corpus-specific models for OCR.

Transkribus was initially designed to decipher manuscripts [4]. It allows the manual transcription of uploaded documents so that they can be used as training material for **Handwritten Text Recognition** (HTR) models (Weidemann et al., 2017). A useful feature of Transkribus' HTR models is that the recognition of printed texts works just as well as that of manuscripts. A few dozen of corrected pages are sufficient for high-quality OCR results.

In this study we illustrate how to drastically improve OCR quality for black letter in newspapers with a modest amount of manual work for ground truth creation. The integration of HTR model training into the Transkribus platform enables Digital Humanists to leverage the performance of neural OCR without having to tackle unnecessary technicalities. In our experiments we additionally address the following questions. Robustness: Are HTR models reusable for material that varies in digitisation quality (medium-resolution scans from microfilm vs. high-resolution scans from paper). Transferability: How well does a model perform on another newspaper than the one it was trained on?

2. Data and Experiments

We use PDFs with medium-resolution images produced in 2005 from scanned microfilms of the German-language **Neue Zürcher Zeitung** (NZZ) for our experiments. The OCRed text stems from **ABBYY FineReader XIX** [5], which was ABBYY's product for 19th century black letter recognition at that time.

The first experiment evaluates the differences between three OCR systems: (a) FineReader XIX (FRXIX) results from 2005, (b) ABBYY FineReader Server 11 (FRS11) results [6], (c) Transkribus' HTR model. Figure 1 shows

example output from our three OCR systems.

Figure 1. Example excerpts with low-quality OCR from two pages of the NZZ (1819 left, 1859 right, red: FRXIX, blue: FRS11, green: Transkribus HTR)

In our second experiment we apply the HTR model trained on medium-resolution images to high-resolution images (400dpi) from 1899 digitised anew from paper in order to test the transferability of the model. We also analyse the performance of the HTR model in two other publications.

2.1. Creation of a ground truth and HTR model training

The NZZ had been published in black letter from 1780 until 1947. We chose one title page per year at random from this period and loaded the image extracted from the PDF into Transkribus. We used the Transkribus internal FRS11 to recognise the text in the images and manually corrected words and baselines. The resulting ground truth of 167 pages contains 304,286 words and 43,151 lines. Depending on the amount of text on a page, the correction of a page including the baselines (needed to train the HTR model) takes between 1 and 2.5 hours. We used 90/10 split for training and testing the model.

2.2. Evaluation

We use the bag-of-words F1-measure metrics of PRImA TextEval 1.4 [7] for evaluation. The F1-measure is the harmonic mean of precision and recall. Precision gives the percentage of OCRed words that are part of the ground truth, while recall measures the percentage of ground truth words that were found by the OCR system. By applying a bag-of-words approach, possible differences in layout recognition cannot distort the results.

3. Results

Figure 2 shows the evaluation on all pages from the test set. The FRS11 (mean F1-measure 81.1%, SD 7.3%) beats the FRXIX (mean 67.8%, SD 11.1%) throughout. Our HTR model scores 97.0% (SD, 1.8%) on average and achieves significant improvements over both ABBYY products.

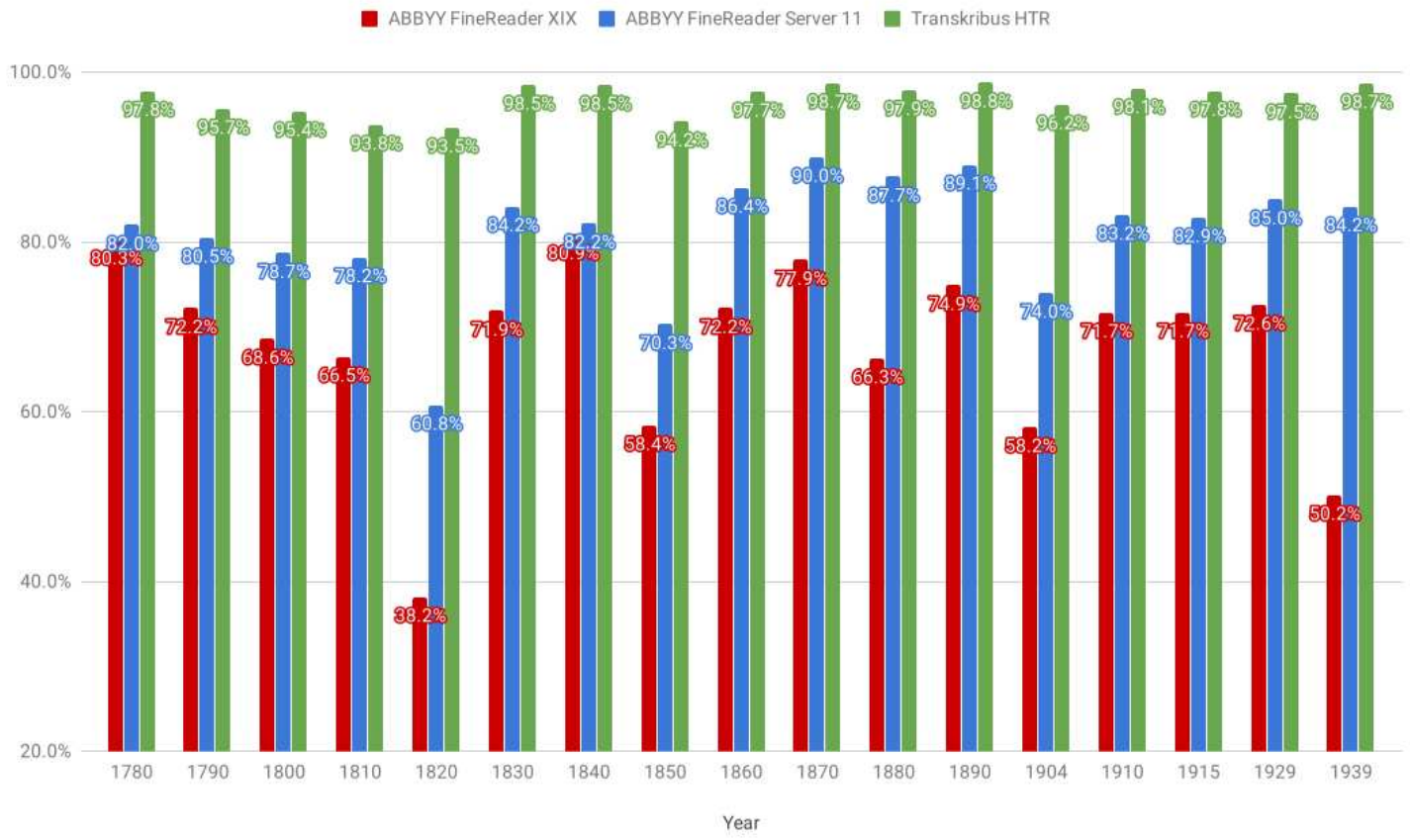

Figure 2. Comparison between original FRXIX, FRS11, and Transkribus HTR.

The application of our HTR model to five high-resolution images of newspaper pages from the NZZ shows accuracies of at least 98% and an average improvement of 4.24% over FRS11 (see Figure 3).

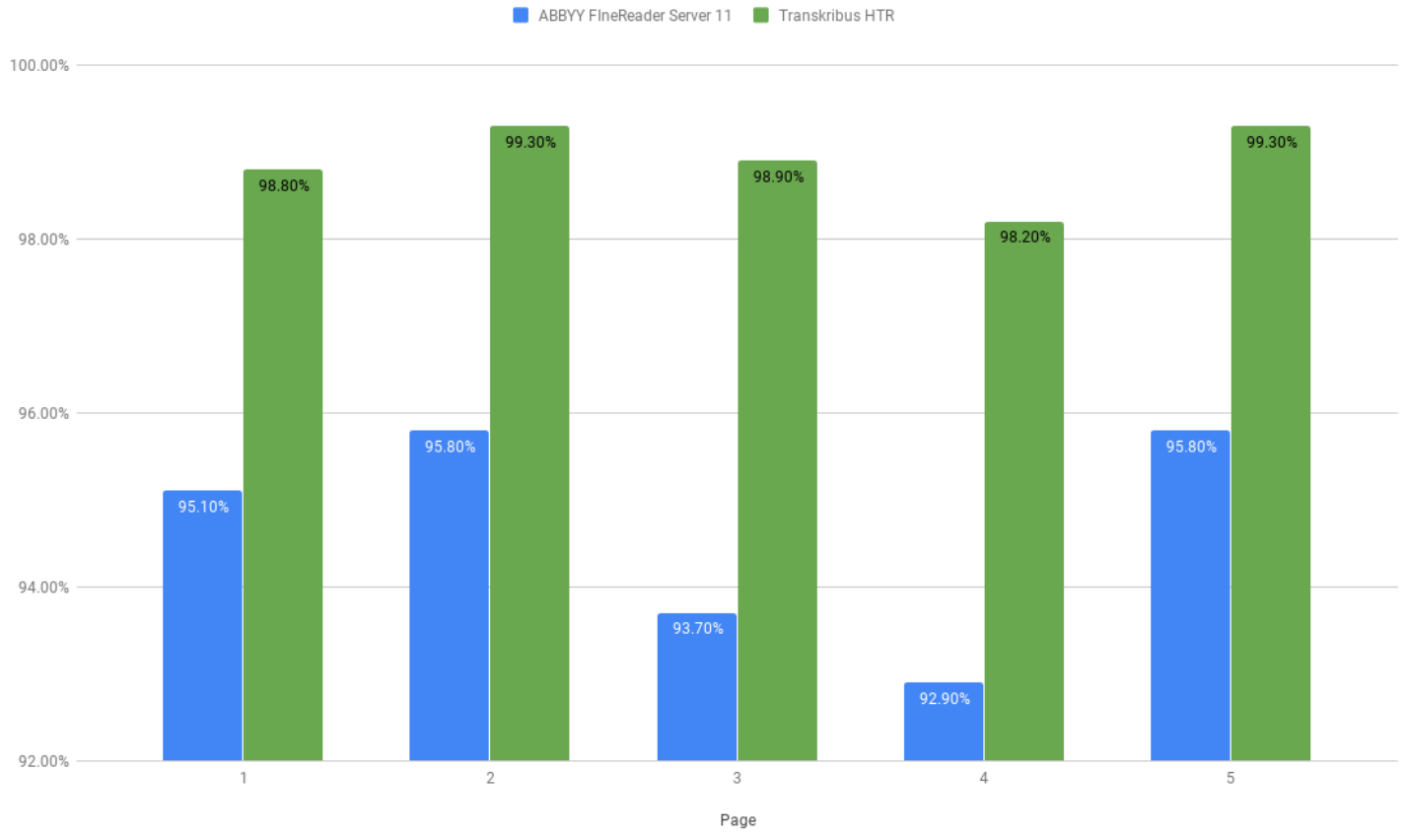

Figure 3. Comparison between FRS11 and Transkribus HTR model on five high-resolution images from 1899.

In terms of the transferability of our HTR model the average F1-measures of 98.6% (SD 1.9%) for the **Bundesblatt** and 98.9% (SD 0.6%) for the **Neue Zuger Zeitung** over five pages each show that although the model has been trained on the NZZ, it is able to score equally high on different publications. The FRS11 reaches 92.4% (SD 2.7%) for the Bundesblatt and 88.4% (SD 3.7%) for the Neue Zuger Zeitung, showing the superiority of our HTR model.

4. Conclusion

We have shown that Transkribus is an excellent tool for creating HTR models for the OCR of newspapers typeset in black letter. Even with a limited amount of training data (150 pages), our HTR model consistently outperforms state-of-the-art commercial software. Our HTR model trained on medium-resolution images digitised from microfilm still performs better than commercial software when applied to high-resolution images derived from paper originals.

Given the availability and abundance of digitised historical material in the form of PDF files with poorly OCRed text, our findings showcase how digital humanists can improve their source material for text mining with a reasonable effort.

5. Acknowledgments

We would like to express our gratitude to Günter Mühlberger and the Transkribus team for their support in training HTR models and partially correcting baselines of our ground truth. Moreover, we thank Camille Watter and Isabel Meraner for their help in the transcription process. This research is supported by the Swiss National Science Foundation under grant CR-SII5_173719.

Appendix A

Bibliography

1. **Alex, B., Grover, C., Klein, E., Tobin, R.** (2012). *Digitised Historical Text: Does it have to be mediOCRe?*, in: KONVENS. pp. 401–409.
2. **Chiron, G., Doucet, A., Coustaty, M., Visani, M., Moreux, J.-P.** (2017). *Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information. 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE. https://doi.org/10.1109/jcdl.2017.7991582
3. **Holley, R.** (2009). *How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs*. D-Lib Magazine 15. https://doi.org/10.1045/march2009-holley
4. **Springmann, U., Lüdeling, A.** (2016). *OCR of Historical Printings with an Application to Building Diachronic Corpora: A Case Study Using the RIDGES Herbal Corpus*. CoRR abs/1608.02153.
5. **Strange, C., McNamara, D., Wodak, J., Wood, I.** (2014). Mining for the Meanings of a Murder: The Impact of OCR Quality on the Use of Digitized Historical Newspapers. *DHQ: Digital Humanities Quarterly 8*.
6. **Walker, D. D., Lund, W. B., Ringger, E. K.** (2010). *Evaluating Models of Latent Document Semantics in the Presence of OCR Errors. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 240–250.
7. **Weidemann, M., Michael, J., Grüning, T., Labahn, R.** (2017). *HTR Engine Based on NNs P2 Building Deep Architectures with TensorFlow*. https://read.transkribus.eu/wp-content/uploads/2017/12/Del_D7_8.pdf

Notes

1.
https://www.abbyy.com
2.
https://github.com/tesseract-ocr/tesseract
3.
https://www.transkribus.eu
4.
see https://read.transkribus.eu/
5.
https://www.frakturschrift.com/de:products:finereaderxix

6.
see https://www.abbyy.com/de-de/finereader-server/ , available within Transkribus
7.
https://www.primaresearch.org/tools/PerformanceEvaluation
